DOCUMENT RESUME

ED 334 209 TM 016 626

AUTHOR Schael, Jocelyne; Dionne, Jean-Paul
TITLE Judges' Agreement and Disagreement Patterns When Encoding Verbal Protocols.
PUB DATE Apr 91
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; *College Graduates; Comparative Analysis; *Encoding (Psychology); *Evaluators; *Interrater Reliability; Nonverbal Communication; *Protocol Analysis; *Sex Differences; Verbal Communication
IDENTIFIERS *Verbal Routines

ABSTRACT

The basis of agreement or disagreement among judges/evaluators when applying a coding scheme to concurrent verbal protocols was studied. The sample included 20 university graduates, from varied backgrounds; 10 subjects had and 10 subjects did not have experience in protocol analysis. The total sample was divided into four balanced groups according to gender and experience in protocol analysis. Subjects were asked to think aloud while encoding two protocols--extracted from an ill-designed task--with a grid of categories provided to them. Subjects' verbalizations and non-verbal behaviors and manipulations were recorded. The level of agreement on the final decisions was analyzed under several criteria. The overall agreement among subjects on the selected codes was 61.4%, but the agreement on the reasons for choosing these codes was 85.1%. This marked difference indicates that the codes selected by the evaluators were not always supported by the reasons stated or that the grid categories are ambiguous. The results show that experience with the task presented (i.e., planning a three-course meal) seemed to be more important for agreement than experience with the technique of protocol analysis. Criteria for improving the quality of a grid and the reliability of coders are suggested. Three data tables and two bar graphs are included. The two study protocols and the coding grid are presented. (Author/RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334209

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JOCELYNE SCHAEL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Judges' Agreement and Disagreement Patterns When Encoding Verbal Protocols

Jocelyne Schael
Jean-Paul Dionne
University of Ottawa, Ontario

Paper presented at the Annual Meeting of the
American Educational Research Association
Chicago, Illinois, April 5, 1991

TM016626

BEST COPY AVAILABLE

Judges' Agreement and Disagreement Patterns When Encoding Verbal Protocols

by J. Schael & J.-P. Dionne
University of Ottawa, Canada

Jocelyne Schael
University of Ottawa
Box #82
145 Jean-Jacques Lussier
Ottawa, Canada, K1N 6N5

Abstract

The present research investigates the basis of agreement or disagreement among judges when applying a coding scheme to concurrent verbal protocols. The sample consists of 20 university graduates, from varied backgrounds, 10 with experience in protocol analysis and 10 without, each subsample balanced for gender. Judges were asked to "think-aloud" while encoding protocols with a grid of categories provided to them. The level of agreement on the final decisions was analyzed under several criteria. The overall agreement among judges on the selected codes was 61.4% but the agreement on the reasons for choosing those codes was 85.1%. The marked difference indicates that the codes selected by the judges are not always coherent with the reasons invoqued or that the grid categories are ambiguous. The results of the study showed that the experience with the task presented (planning a three course meal) seemed to be more important for agreement than experience with the technique of protocol analysis. Criteria to improve the quality of a grid and the reliability of coders are suggested.

Objectives

The increased emphasis in the use of verbal reports as data to study cognitive processes and structures has raised the issue of reliability of encoding verbal protocols. In the present study, reliability characteristics are obtained through the identification of agreement/disagreement patterns among judges performing an encoding task. More precisely, the objectives are:

1. To estimate the level of agreement/disagreement among judges with varied backgrounds (level of experience, gender) performing an encoding task.

2. To contrast the selection of categories by the judges with the reasons invoked for their decisions.

Theoretical framework

The use of verbal reports as data has proven successful in many fields. The technique of protocol analysis, in exploring cognitive structures and processes, evolved with the information processing model of the human mind. Of particular interest, if one is concerned with obtaining concurrent verbal reports, subjects are asked to "think-aloud" while they perform a task. Thus, subjects pursue two tasks concurrently: the verbalization of their thoughts and the execution of the task. To analyze the recorded verbalizations, the experimenter identifies categories of behavior (coding scheme) based on a theoretical model and data grounded. Many researchers in the areas of memory and problem solving, in particular Newell and Simon (1972), have used the technique extensively.

The interpretation of protocols within an information processing framework generally relies on three fundamental assumptions: one can describe a cognitive process by the information that is attended to (heeded) in short term memory (STM) as the stimulus is encountered and processing takes place; attending to information takes time; and, the capacity of STM is limited. These assumptions form the basis for analyzing the cognitive trace through concurrent verbal protocols.

Ericsson and Simon (1984) have elaborated on the technique of protocol analysis, providing extensive information on the procedure involved. From the research literature, reliability of encoding has been assessed with well-defined problems and has proven to be high (Ericsson & Simon, 1984). However, little has been developed for testing the reliability of encoding complex processes from performances on ill-defined tasks. Few and non systematic attempts have been made: they are usually done using two to five raters; minimum information is

provided as to how the inter-coder reliabilities are measured (Goor & Sommerfeld, 1974; Haines, 1974; Klinger, 1974; Svenson, 1983). As an extension of previous work, the present research is based on protocols extracted from an ill defined-task and collected from a large number of raters (20) of varied background (level of expertise, gender).

Methodology

University graduates and professors represent the pool of candidates from which judges are likely to be selected for this type of research: assessing the reliability of information extracted from verbal protocols.

Sample

It was expected that the experience of judges at encoding could influence the reliability of the technique. To test the influence of experience with encoding, ten judges with experience in protocol analysis and ten judges without such experience formed the sample. It is also possible (Ericsson & Simon, 1984) that judges may encode with a bias toward their own preferred interpretation. To alleviate this problem and reduce possible bias, judges with varied academic backgrounds were selected. Due to the nature of the task (encoding protocols collected while planning a three course meal), gender may have had an effect on the reliability level of the results. The sample was therefore balanced for gender. The total sample of 20 graduate students and professors from varied fields of specialization was divided into four balanced groups according to experience in protocol analysis and gender.

Task

Judges were presented with three protocols (S1, S2, and S3) and the coding scheme (grid) developed in the Gillespie (1988) study. In the training phase of the experiment, judges were asked to "think-aloud" while encoding Protocol S1; this gave them the opportunity to familiarize themselves with the "thinking-aloud" technique as well as the task. The main task consisted in encoding Protocols S2 and S3. Hereafter, **S-protocols** refer to the two protocols to be encoded and **J-protocols**, the ones obtained from the judges in the present study. The procedure can then be summarized as the analysis of **J-protocols** while the judges encode **S-protocols.**

The **S-protocols** were collected from students in a College of Agriculture and Technology who were asked to "think-aloud" while performing a curriculum task, the planning of a three-course meal for six persons. A feature of protocol S2 is that it was taken from an expert at the College, while the other (protocol S3) was collected from a freshman. The complete protocols (S2 and S3) are in the Appendix.

The grid used was generated from the analysis of the **S-protocols** (Gillespie, 1988). Thirty-nine mental activities performed by the students while solving the menu planning task were extracted from their protocols. These operations were grouped under 7 broad categories:

P: Problem identification and representation

S: Solution related activities

M: Memory related activities

C: Changing the conditions of the problem

R: Reasons for particular food choices

A: Assessing and monitoring activities

E: Emotional reactions.

The complete grid is in the Appendix. Judges were allowed to assign multiple codes and were asked to write their final answers next to each statement. They were instructed to verbalize the reasons leading to their decisions.

Data

The experimenter recorded the judges' verbalizations along with observations of non-verbal behaviors, manipulations, interruptions and time lapses. The tapes and observations were later transcribed. The codes selected by the judges (final decision) and their justification for each decision (verbal J-protocols) provided experimental data. Both quantitative (code frequencies) and qualitative (verbal protocols) data were extracted from the judges responses. The two sets of data yielded information on the congruence between the selection of categories (codes) by judges and the reasons given to arrive at that decision. Comparisons among different groups provided information on patterns of agreement/disagreement.

Results and conclusion

The results showed a wide range of agreement among judges over categories. The level of agreement on the final decisions was analyzed under several criteria. When calculated from the codes selected by the judges the overall agreement among judges was 61.4% (54.1 + 68.6, for S2 and S3) but when calculated from the reasons invoked for these codes, i.e., the verbalized responses collected during the execution of the encoding task, the agreement was

85.1% (81.0 + 89.0, for S2 and S3). These results are presented in Table 1 in the Appendix.

The results were first interpreted in relation to the selected variables. In a second analysis the results from the written indices are compared to the results from the verbal indices. Finally, differently selected agreement/disagreement patterns are interpreted.

Results considering the selected parameters. When the four groups are examined separately, the levels of agreement on the written indices are higher than those for the entire group of 20 judges (Table 1 in Appendix). It was observed that the code on which the smaller group agreed is sometimes different than the code chosen by the majority in a larger group. This would support Ericsson and Simon's (1984) claim that a small group of coders may decide on a code with a bias toward their own preferred interpretation which could be different from another group with different views. From these results, special attention should be given to the number of judges selected for a reliability check - when a small number of judges from homogeneous backgrounds is selected, the agreement level may be inflated and biased towards the selected group.

An examination across **S-protocols** showed a greater level of agreement for all groups on the protocol of the freshman at the College, Protocol S3 (Table 1 in Appendix). The highest indices of agreement were found on both **S-protocols** when the link between the defined category and the students' verbalizations was direct and the level of inference was low. Disagreement was high when the context of the task was not fully appreciated by the judges.

The results were then examined according to the experience variable. No significant association between experience and level of agreement was found for

either the written index (χ^2 = 6.88, df = 3, p = .0750) or the verbal index (χ^2 = 2.69, df = 3, p = .5563). An indication that can account for these results presented in Table 2 of the Appendix, is the richness of the "no experience" judges' protocols; richness in this instance meaning the quality as well as the quantity of the verbalizations. Their respective protocols could be compared to novices protocols which often contain every step of a solution, as opposed to the protocols of experts in which processes would have become condensed reports.

A significant association between gender and level of agreement was observed for both the written index (χ^2 = 10.59, df = 3, p = .0144) and the verbal index (χ^2 = 8.24, df = 3, p = .0410). These results are presented in Table 3 of the Appendix. From the analysis of the statements on which the agreement level was higher for women than for men, women in this sample showed more "expertise in the field of cooking". They displayed a richer knowledge base relative to specific methods and technical terminology. It was concluded that the level of task dependent expertise seemed to have more of an impact on agreement among judges than experience with the technique of protocol analysis.

Comparison of written and verbal indices. Results from the comparison analysis (Figures 1 & 2 in Appendix) showed higher level of agreement when verbalized justifications for the selected codes were taken into consideration. The marked difference indicated that the codes selected by the judges were not always coherent with the reasons invoked or that the grid categories were ambiguous. More specifically,

- the wording in the definitions of the codes may have had many meanings, thus permitting two or more interpretations to be coded differently;
- judges had very little time to familiarize themselves with the numerous categories;

- some categories seemed to be missing, incomplete, or too general.

These findings can be summarized as being important indicators for increasing reliability of encoding verbal protocols: preparation of the judges prior to encoding, and elaboration of complete and discriminating categories. Preparation of the judges is reduced when prototype examples are used in the definitions of the categories. Complete and discriminating categories are easier to develop when the theoretical model is grounded into the data. With this information it may be argued that the level of agreement using the written codes, would increase substantially.

Ageement/disagreement patterns according to judges' characteristics. Comparison of agreement/disagreement was made between judges who tried to analyze each segment independently and judges who chose to analyze by larger sections. The main findings indicated two sources of problem which could affect the reliability of encoding. The first one came from the group of judges who analyzed segment independently: total independence of segments was not always possible, information from other statement was sometimes necessary for accurate interpretation when encoding protocols collected while performing an ill-defined task. The second problem was encountered in the data collected from the judges who analyzed by larger sections: using context, decisions were not always independent of each other.

Two main characteristics emerged from the observations collected from the judges: (1) task control judges, the ones who were concerned with choosing a description which would match the statement - finding the correct answer on the grid; (2) and the self control judges who were concerned with accurate interpretations of statements, chosing a code for which they had found good arguments to support their choice. Agreement/disagreement indices were

compared between "task control" judges and "self control" judges. The results revealed that agreement is generally higher for "self control" judges. A more microscopic study is underway to understand about whether "task control" judges neglect some important information or dispense with making a final decision.

Discussion and Recommendations

This exploratory study provides a basis for determining criteria which could improve the reliability of the encoding technique when verbal protocols are used as data. The findings support the notion that information provided in verbal protocols can reliably be extracted. When a lower consistency was observed in encoding protocols it was caused largely by failure to agree on a specific category, not in interpreting verbal information. Criteria for developing a reliable grid, choosing judges, and designing an encoding task are suggested from the findings.

Developing a reliable grid:

- a grid must contain a category for every episode to be encoded (integrate data based grid development);
- generality or specificity of definitions should be maintained throughout the grid (unilevel categories);
- definition of categories must discriminate well (use prototype examples);
- the number of codes should be limited.

Choosing judges:

- the knowledge base is a relevant factor;

- judges providing reasons for their choices seemed to characterize the more reliable judge;

- if cognitive processes used by judges when encoding verbal protocols are being analyzed, a person with less experience in protocol analysis may provide more information than a person with extensive experience with the technique.

Designing an encoding task.

An ill-defined problem can be encoded with a high level of agreement as the results from the verbal indices indicated.

- Protocol segmentation seemed to have an influence on the justification for a choice, if each segment refers to a unique code the verbalized justifications of judges would be related to that idea.

- Open discussion between judges and experimenters following the practice session is of prime importance for reducing various semantic problems.

- Practice can also include discussion on the type of information to be encoded: (1) what the subject is doing, and (2) what the subject is thinking.

- Information about the physical environment in which the protocol was produced is important for the judges encoding protocols because it is sometimes difficult to capture adequately but essential to reliable encoding.

- Segments used in context do not provide evidence which is wholly independent from other segments. The amount of context that must be considered when encoding a segment is a function of the type of encoding: low-level encoding or high-level encoding.

Educational implication

The results of the present study have helped established criteria to improve the reliability of a process tracing technique. Educators are placing more emphasis upon the understanding of processes underlying performance in learning situations. Reliable identification of processes within a problem-solving context can benefit instruction and learning in the educational system.

Concurrent verbal reports provide researchers and educational practitioners with insights about what it is students are doing while they seek to understand and learn. More information on cognitive processes will help teachers guide students in their thinking. Integration of reliable process data and product oriented data will give a more comprehensive view of intellectual abilities.

Selected references

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, Massachusetts: The MIT Press.

Gillespie, S. (1988). [Verbal protocols collected at a College of Agriculture and Technology]. Unpublished data.

Goor A., & Sommerfeld, R. E. (1974). A comparison of problem-solving processes of creative students and noncreative students. Journal of educational Psychology, 67, 495-505.

Haines, G. H. (1974). Process models of consumer decision making. In G. D. Hughes & M. L. Ray (Ed.), Buyer/consumer information processing. Chapel Hill: University of North Carolina Press.

Klinger, E. (1974). Utterances to evaluate steps and control attention distinguish operant from respondent thought while thinking out loud. Bulletin of the Psychonomic Society. 4, 44-45.

Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Svenson, O. (1983). Scaling evaluative statements in verbal protocols from decision processes. In P. Humphreys, O. Svenson, & A. Vari (Ed.), Analysing and aiding decision process. Amsterdam: North Holland.

Appendix

Protocol S2

my food list

well I'll have a cup of coffee first

now my food list the rest of my food list

spices and alcoholic beverages

my time sheet for preparing my menu

my sheet for writing down my menu

ok now food list and a menu sheet and a time sheet

design a three course menu

check off those food items you plan to use

so I pretty well have to know the recipes for what I'm using

so I could either start with the soup

or a salad

I wonder if there's any avocados on this

let's see no no avocadoes

how about onions no only cooking onions

there's mushrooms

well I guess we should go for something classic

E: tell me what you're writing as you go

S: ok French onion soup

and I check off my cooking onions

and I'm going to need something to sauté those in

so let me see do you have any shortening lard or vegetable oil

oh I better be a good dietitian here and use vegetable oil it

Protocol S2 (continued)

doesn't have as much saturated fat

and a little bit of flour and a little bit of sugar

and a little bit of salt where's the salt?

E: it's on the second sheet

S: ok there's my salt and then I want some beef some consommé

ok now I need some bread for some croutons hope I don't have to make the bread

here we are here's some fresh bread so that maybe we'll just put a slice of French bread instead of the croutons

and some cheeses somewhere

what do we have

we have Mozzarella, and Cheddar, and Brie

well can't grate Brie very well so I'll use the Mozzarella and I guess I don't have any Gruyère or Swiss cheese

well we'll mix it with Cheddar give it a little bit more flavour

so that takes care of the onion soup

and ok lamb chops are nice we could have broiled lamb chops I'll write that broiled lamb chops

and our vegetable here broiled lamb chops ok with broccoli

and we could put some rice and mushrooms with that so

ok mushrooms we need here's our rice need a bit of butter with the broccoli I don't see any butter

E: it's in the dairy section

S: the dairy products ok

ok that takes care of our broiled lamb chops, rice and mushrooms,

and some broccoli

and now I suppose if I serve a salad with that I've lost my three

Protocol S2 (continued)

courses

ok so we need a dessert

how about a baked alaska that sounds nice

ice cream and eggs whites and sugar

I have sugar checked already so

baked alaska, now that's all and I suppose tea or coffee

what about bread and butter is that considered a add that as well

so maybe some nice rolls and I'll buy those so I don't have to make them instant dinner rolls and I already have my butter

and then a beverage

tea coffee or milk and I suppose we want a red wine with this

ok so that's my menu

Protocol S3

I'm taking the food list and I'm going to pick out the foods
I'd like for this dinner for six people or six adult

um we'll have coffee

do I have to write the amount
(yes)

um 15 no 2000 ml. of coffee in the pot and um

we're going to have some um some chicken breasts

have 12 is this where I write them down (yes)

um oh um I'll have carrots

um 500 ml. um

dinner rolls 12 dinner rolls and

some potatoes 6 potatoes and

for dessert um oh um what goes with a three course meal

um cream of chicken soup 6 cans I guess

and um long pause ice cream for dessert

1500 ml.

E: could you speak out a little bit louder

S: oh ok um ok oh I'll have take the pantry stock or page two

E: yes that goes with your food list

S: oh um we're going to have dry red wine no chicken yes

2 bottles

then I'll um make up the time sheet

Coding Grid

LOCAL PLANNING OPERATIONS

Subject uses a combination of these operations

P: Problem Identification and Representation

- P1 - Reads instructions
- P2 - Rereads or summarizes all or part of instructions
- P3 - Identifies or defines problem
- P4 - Chooses strategy for attacking the problem
- P5 - Refers to an element in the problem statement
- P6 - Questions E about problem (for clarification only)
- P7 - Fails to consider all important elements in problem
- P8 - Does not have specific vocabulary/domain specific knowledge to understand problem. Is confused, therefore, about the task

S: Solution Related Activity

- S1 - Designs appetizer - that course which comes before the main part of the meal; may consist of soup(s), salad(sl), hors d'oeuvres(h), beverage(b)
- S2 - Designs the main course of the meal; may consist of fish/meat(m), potatoes/pasta/rice(c), vegetable(v), salad(sl), beverage(b)
- S3 - Designs dessert; may consist of food(f), beverage(b)
- S4 - Garnishes a particular dish for aesthetic appeal
- S5 - Lists ingredients in recipe for a particular dish
- S6 - Interrupts work on one course/dish to go back to elaborate on an already designed component
- S7 - Interrupts work on time-planning task to go back to menu-planning task and make additions/changes to it
- S8 - Substitutes one food item for another to adhere to restrictions of the food lists
- S9 - Modifies or changes a dish for reasons other than #S8
- S10 - Calculates amounts/quantities of food items needed for six people
- S11 - Calculates cost of a particular item
- S12 - Works on time-planning task, deciding when and how to prepare foods for serving
- S13 - Misjudges amounts appropriate for six people
- S14 - Confuses amount of food/beverage served with amount of raw ingredient needed to prepare the dish

Coding Grid (continued)

M: Memory Related Activity
- M1 - Attempts to recall previous experience to use as a model
- M2 - Attempts to associate, compare, relate to previous experience outside task

C: Changing the Conditions of the Problem
- C1 - Requests that new items be added to the food lists
- C2 - Consciously uses items which are not on the food lists

R: Reasons for Particular Food Choices
- R1 - Explains that a food choice has been made for aesthetic reasons, relating to sensory appeal
- R2 - Explains that a food choice has been made for nutritional reasons
- R3 - Indicates awareness of cost factor when choosing a particular food
- R4 - Explains that a food choice has been made on basis of cost
- R5 - Explains that a food choice has been made to illustrate subject's expertise in the field

A: Assessing and Monitoring Activity
- A1 - Locates position with respect to entire task (monitoring)
- A2 - Summarizes what has been done
- A3 - Assesses a portion of the task
- A4 - Makes changes/additions to the menu as a result of the assessment process
- A5 - Evaluates the entire menu at the end of the task

E: Emotional Reactions
- E1 - Comments explicitly on the difficulty (-) or ease (+) of problem
- E2 - Expresses feeling of confusion or frustration with the task
- E3 - Expresses feeling of accomplishment, relief, or pleasure from the task

Table 1

Mean Percent Level of Agreement for Selected Groups of Judges

Groups	n	Written indices		Verbal indices	
		Protocol			
		S2	S3	S2	S3
Experienced					
Male	5	56.0	71.1	79.6	85.6
Female	5	63.1	71.1	78.7	86.7
Non-experienced					
Male	5	60.4	75.6	77.8	87.8
Female	5	70.7	74.4	88.0	88.9
Total group	20	54.1	68.6	81.0	89.0

Table 2

Chi-square Results for Experience Variable

	Number of judges agreeing						
Experience	2	3	4	5	χ^2	df	p
				Written index			
With	45	23	29	23			
Without	28	34	36	25	6.88	3	.0750
				Verbal index			
With	11	17	36	50			
Without	7	13	41	62	2.69	3	.5563

Table 3

Chi-square Results for Gender Variable

	Number of judges agreeing						
Gender	2	3	4	5	χ^2	df	p
				Written index			
Male	47	31	25	21			
Female	26	26	40	27	10.59	3	.0144
				Verbal index			
Male	13	18	39	46			
Female	5	12	38	66	8.24	3	.0410

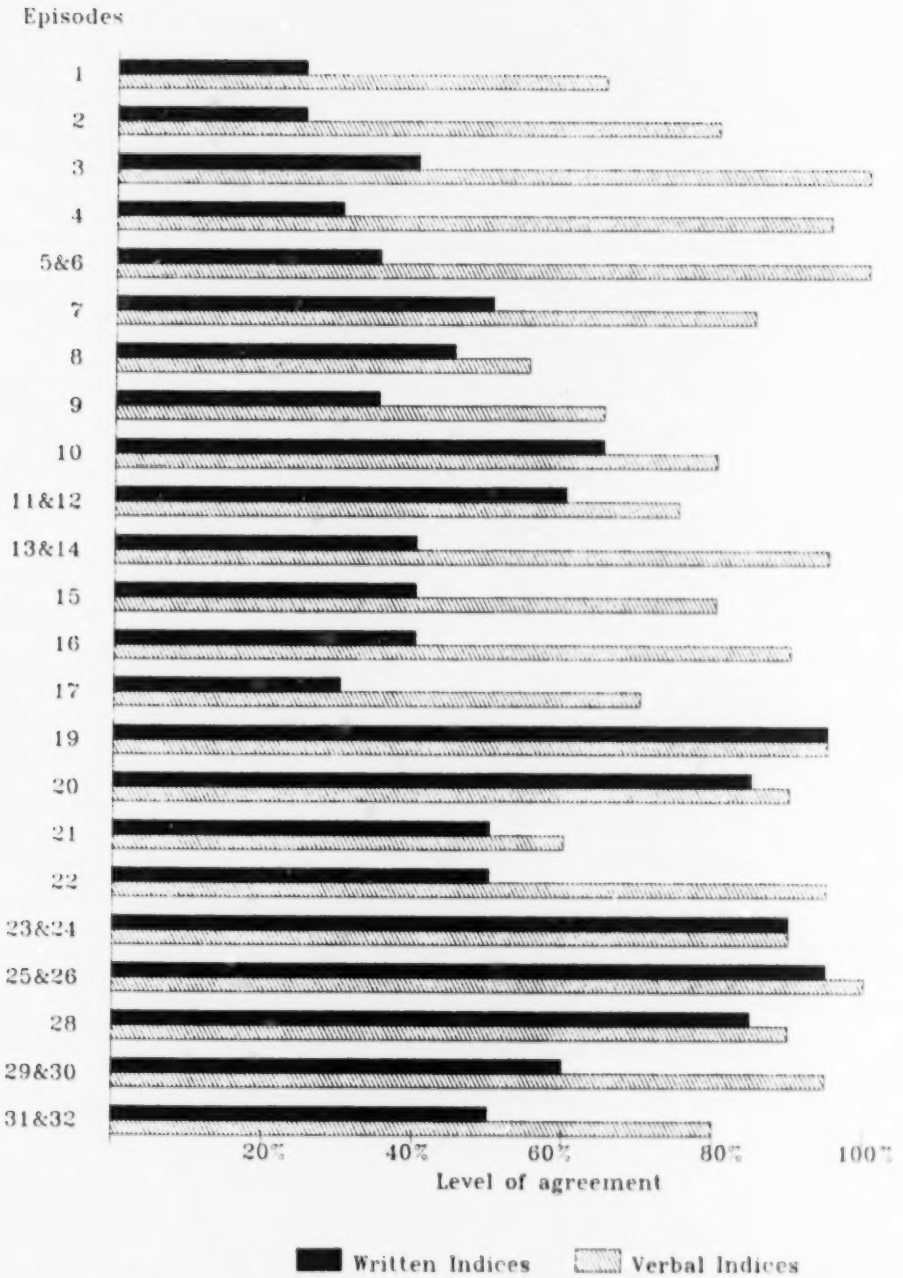

Figure 1. Bar graph representation of verbal and written indices of judge agreement on episodes of protocol 2.

BEST COPY AVAILABLE

Figure 1 (continued)

Figure 2. Bar graph representation of verbal and written indices of judge agreement on episodes of protocol 3.

BEST COPY AVAILABLE